BMC Medical Informatics
and Decision Making

RESEARCH Open Access

Predicting the risk of diabetes complications using machine learning and social administrative data in a country with ethnic inequities in health: Aotearoa New Zealand

Nhung Nghiem[1,2*], Nick Wilson[1], Jeremy Krebs[3] and Truyen Tran[4]

Abstract

Background In the age of big data, linked social and administrative health data in combination with machine learning (ML) is being increasingly used to improve prediction in chronic disease, e.g., cardiovascular diseases (CVD). In this study we aimed to apply ML methods on extensive national-level health and social administrative datasets to assess the utility of these for predicting future diabetes complications, including by ethnicity.

Methods Five ML models were used to predict CVD events among all people with known diabetes in the population of New Zealand, utilizing nationwide individual-level administrative data.

Results The Xgboost ML model had the best predictive power for predicting CVD events three years into the future among the population with diabetes (*N* = 145,600). The optimization procedure also found limited improvement in prediction by ethnicity (using area under the receiver operating curve, [AUC]). The results indicated no trade-off between model predictive performance and equity gap of prediction by ethnicity (that is improving model prediction and reducing performance gaps by ethnicity can be achieved simultaneously). The list of variables of importance was different among different models/ethnic groups, for example: age, deprivation (neighborhood-level), having had a hospitalization event, and the number of years living with diabetes.

Discussion and conclusions We provide further evidence that ML with administrative health data can be used for meaningful future prediction of health outcomes. As such, it could be utilized to inform health planning and healthcare resource allocation for diabetes management and the prevention of CVD events. Our results may suggest limited scope for developing prediction models by ethnic group and that the major ways to reduce inequitable health outcomes is probably via improved delivery of prevention and management to those groups with diabetes at highest need.

Keywords Machine learning, Diabetes complications, Cardiovascular disease, Risk prediction, Health and social administrative data

*Correspondence:
Nhung Nghiem
nhung.nghiem@anu.edu.au
[1] Department of Public Health, University of Otago Wellington, Wellington City, Wellington 6021, New Zealand
[2] John Curtin School of Medical Research, Australian National University, Canberra City, ACT 2601, Australia
[3] Department of Medicine, University of Otago Wellington, Wellington City, Wellington 6021, New Zealand
[4] Applied Artificial Intelligence Institute (A2I2), Deakin University, Geelong City, VIC 3216, Australia

Background

People living with diabetes have a higher risk for cardiovascular disease (CVD) events than the general population [1]. According to the Global Burden of Disease Study 2017, CVD is the leading cause of death in the world [2]. Some treatments for CVD can be very expensive and cumulatively account for a large proportion of total health system costs [3, 4]. Therefore predicting CVD events among people with diabetes is desirable for health system planning. In addition, diabetes and CVD events are more prevalent in some ethnic groups than the others [5, 6], and this needs to be taken into account in health outcome prediction. In Aotearoa New Zealand (NZ), diabetes and CVD are the leading causes of premature death and disease burden, and are major sources of health inequities for Māori, Pasifika, and Asian populations due to socio-economic, cultural and health system factors. For example, Māori are 3.13 times more likely to suffer diabetes complications and 1.46 times more likely to have moderate CVD compared to the European/other group [7]. Therefore, improving risk prediction for CVD could provide opportunities to improve the health and lifespan of individuals and ethnic groups.

There is strong evidence that controlling glucose levels and hypertension, managing dyslipidemia, and smoking cessation can reduce the risk of people with diabetes developing CVD [8]. However, there are factors at a system-level, which compromise the ability to intervene upon this evidence and care for populations, such as socio-economic status, medication costs, and access to healthcare [9]. Health inequities in NZ have long been recognized, yet little improvement has been achieved over the last 20 years or more [10].

The NZ Government, similar to the governments in Scandinavian countries, United Kingdom, Canada and Australia, holds a large amount of data from patient interactions with the healthcare system [11]. This is in addition to extensive other individual data such as census, immigration, and justice data, which can be linked at an individual level. These data are high-dimensional, very extensive and impossible to explore by clinicians or health systems decision makers manually.

Machine learning (ML) method has emerged as a promising new technique to model disease risk prediction in an era of big datasets [12–15]. It consists of a large number of alternative methods including classification trees, random forest, neural networks, support vector machines, and lasso and ridge regression. For studies where the primary goal is to predict the occurrence of an event, this technique produces a more flexible relationship among the predictor variables and the outcome [12]. The advantages of ML over the traditional regression models are that ML can handle non-linear relationships efficiently without any specification of the functional form that links the model's features and the predicted outcome [16–18]. ML is also suitable for handling high-dimensional and large datasets [19]. In this study, traditional regressions refer to both linear and non-linear regression models. However, traditional regression models are central around parameter estimations with specified functional forms, while ML focuses on outcome predictions [20]. In fact, the emerging evidence suggests that ML significantly improves accuracy of CVD risk prediction compared to the traditional regression models [9, 16, 18, 21].

There are a large number of prediction models that have been developed for CVD events among people with diabetes in the clinical setting [9], including both traditional regression and ML methods [1, 22]. These models generally utilize rich clinical information or features (e.g., body mass index, smoking status, biomarkers) extracted from electronic medical records or clinical trials. However, while these models are important for risk prediction at a clinical level, they are not easily deployed at the population level in order to reduce systemic barriers to improve diabetes management. In contrast, linked social and administrative health data consists of records collected on diagnoses, medications, and demographics generated through the provision of health services by governments. These data have become increasingly available to assess population health [11, 23], and they represent a valuable resource for automated analytic approaches to improve the efficiency and effectiveness of primary and secondary health prevention efforts [9].

Given this background, the overall aim of this research was to: (1) use ML models to predict CVD events over a three-year period for the NZ adult population with diabetes using a broad range of routinely collected health and social administrative data; and (2) assess the performance of ML models on different ethnic groups in NZ to determine the relevance to reducing health inequities. We first aimed to determine the best optimization methods based on training datasets to deal with ethnic data, then we used this method to further test the model's performance in predicting future events.

Methods

Datasets

We used linked health and social administrative data from the Stats NZ Integrated Data Infrastructure (IDI). This is a research database that links a broad and diverse collection of administrative and survey datasets from health, income, benefits and social services, education, justice, housing, and communities. All individual data across different datasets were linked through a unique identifier code. Unique identifier codes for individuals were generated and maintained by Stats NZ across different datasets. They first linked individual records from surveys, government agencies, and non-government organizations together using identifiable data, including first and last name, date of birth, age, sex, and country of birth. Then, all these personal identifiers were removed and encrypted to ensure the data confidentiality before they were made available for research. For this study, we obtained access to the datasets of interest, linked individual records using unique identifiers across these datasets, and extracted relevant records.

The first dataset was the Census 2013 to identify individuals' smoking status, language spoken, employment status and other demographic information. The second dataset, the diabetes complications dataset from the Ministry of Health (MoH chronic condition table), contains information about healthcare users in the population cohort who have been diagnosed with one or more of eight chronic conditions (e.g., coronary heart disease, stroke, diabetes, cancer, and gout). We used this dataset to identify people with CVD and diabetes, and other chronic diseases in 2013 [24]. In order to identify individuals on CVD preventive pharmacotherapy, we used pharmaceutical data from 2013, but with no history of a CVD event (i.e., individuals that had: (i) none of the conditions in the MoH chronic conditions table; or (ii) did have one of these conditions, but who had no prior identified CVD condition). This dataset contains claim and payment information from pharmacists for subsidized dispensings. Finally, we used the IDI Population Explorer dataset (2013), which has indicators for receipt of social security benefit, use of social housing, and major life events (i.e., getting divorced/separated when this was officially documented) in 2013 [4]. Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of our research.

Ethical considerations of using large-scale administrative data for health prediction

The large-scale administrative data for health prediction were kept and maintained by StatsNZ in a confidential environment with five relevant principles [25]. This includes safe people; that is, researchers are vetted and must commit to using data safely before they can access the data. Safe projects mean to gain access to integrated data, researchers must have a project they can demonstrate is in the public interest. Safe settings ensure a range of privacy and security arrangements to keep data safe. Safe data means identity is protected. Data has had identifying information removed, and researchers only get access to the data they need. Finally, safe output means that all information is checked to ensure it does not contain any identifying results. For further information with regard to the ethical considerations of using administrative data, please see Declarations sections below: Ethics Approval and consent to participate, Availability of data and materials, and Acknowledgements.

Study population

The whole population of NZ who were in the residential population in 2013, and who had been diagnosed with diabetes in Virtual Diabetes Registry in the period of January 2001 to December 2013 [26] but with no prior CVD, were followed throughout to 2018. In order to identify people with diabetes, any CVD complications they had, and their social characteristics, we used the International Classification of Diseases (ICD) for identifying diabetes and CVD complication events [3, 27]. People with diabetes in the Virtual Diabetes Registry included both Type 1 and Type 2 diabetes (ICD-10-AM: E10, E11, E13, E14, O240-O243). However, there was no validated algorithm to separate these two types of diabetes from this Registry, so we included all these people in our study [26]. In fact, this was consistent with a CVD risk assessment guideline by the NZ Government that included all diabetes types [28]. The definition for CVD was based on the following ICD-10 codes: stroke (I60-I64; G45-G46), and coronary heart disease (ICD-10-AM: I20-I25) [26]. We aimed for forms of CVD that were likely to be associated with diabetes (e.g., excluding valvular disease which is often associated with past rheumatic heart disease in the NZ context), and where there is high specificity to true CVD (e.g., excluding congestive heart failure which can arise from non-CVD causes such as chronic obstructive pulmonary disease).

Only people who were in both the Census 2013 and the IDI estimated resident population in 2013, and had diabetes but did not have diagnosed CVD, were included in the analysis. We also further restricted the population to people aged between 30 and 74 years old as per other NZ work in CVD risk prediction [23]. All observations with missing age and sex information were excluded from the analysis but these were very infrequent. Steps to extract the study population were presented in Fig. 1.

Outcome

This included the risks of developing CVD over a three-year period. The dependent variable was a binary outcome whether a CVD event (either fatal or non-fatal) had or had not occurred for an individual with diabetes during three-year periods between 1/1/2013-31/12/2015; and 1/1/2016-31/12/2018.

Variables

The linked health and administrative datasets allow us to examine the individual-level impact of not only the health indicators (e.g., diabetes, smoking status); but also other demographic characteristics (age, sex, ethnicity [self-identified], immigration status); social background

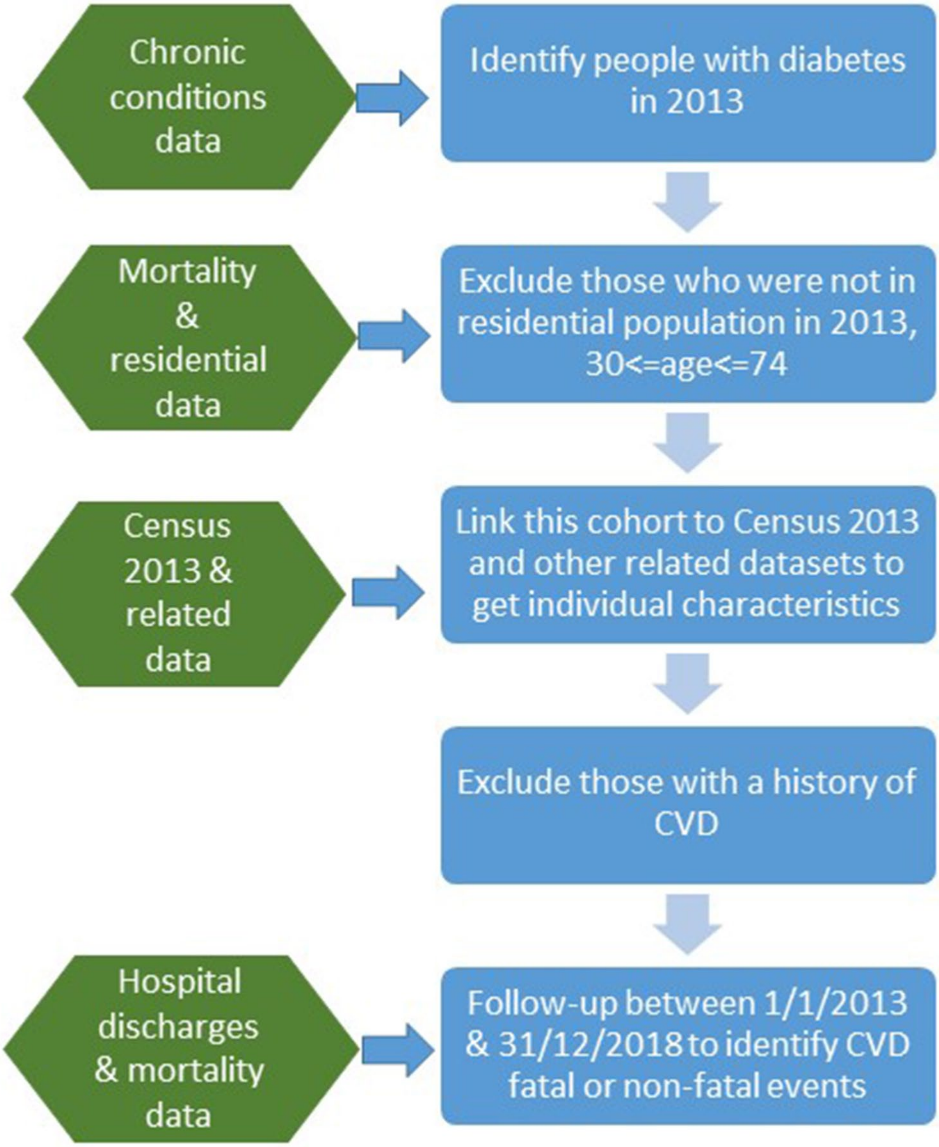
Chronic conditions data
Identify people with diabetes in 2013
Mortality & residential data
Exclude those who were not in residential population in 2013, 30<=age<=74
Census 2013 & related data
Link this cohort to Census 2013 and other related datasets to get individual characteristics
Exclude those with a history of CVD
Hospital discharges & mortality data
Follow-up between 1/1/2013 & 31/12/2018 to identify CVD fatal or non-fatal events

Fig. 1 Steps to extract the study population from the linked administrative and health data

variables (e.g., housing conditions, social security benefits and language spoken); potential stress indicators (via employment); and deprivation quintiles on a one to five scale with five being the most deprived. New Zealand's 2013 deprivation is an official neighbourhood-level measure of socio-economic deprivation, which combined nine variables from the 2013 census survey, reflecting eight dimensions of socio-economic deprivation including internet access, income, unemployment, qualification, single parent, home crowding and access to a car) [29, 30]. We used ethnicity as a proxy for structural socioeconomic factors that affect health because, in the NZ context, diabetes and CVD are manifestations of inequity due to socio-economic, cultural, and health system factors.

In addition, the following conditions were added to the predictor variables: any hospital event between 2001 and 2013 for dementia, asthma, chronic kidney disease, and total hospital events for any condition. Disease ICD-10 codes were extracted from the MoH Burden of Disease Report 2016 (see the Appendix) [31, 32]. These conditions were added to the predictor variable list based on the literature for conditions associated with CVD [33–35].

Data pre-processing

In this research, the whole NZ population (i.e., adults aged 30–74 with diabetes but no prior CVD in 2013) was split into study and validation datasets. Individuals could only be in either the study or the validation dataset. Then, the study dataset was further randomly divided into 80% training and 20% test datasets, using the k-fold principle [36]. Similarly, individuals could only be in either training or test dataset to maintain the validity of the test, that is none of the data involved in fitting the prediction function was used to evaluate the prediction function that was produced [37]. Training involved fitting a model, while testing involved using a subset of data to evaluate empirical performance of a model trained on a training dataset [20, 38]. The final prediction that was developed based on training and test datasets was used to predict CVD event outcomes of individuals in the validation dataset to evaluate the performance of this prediction model against the future events (in the time period: 2016-18). This future prediction can be classified as a form of external validation for the final ML model.

Data subsets by time period and by ethnicity

We split out datasets into a study dataset with a three-year follow-up from 2013 to 2015 and a validation dataset from 2016 to 2018 as above mentioned [9]. We also created datasets by ethnicity within the study and validation datasets, in particular: the whole NZ population with diabetes, Asian population with diabetes, Māori population with diabetes and Pasifika population with diabetes.

Model development and evaluation

We used ML models, including L1-regularized logistic regression, decision trees, random forest, neural network, and Xgboost to predict CVD complications [39–41]. These ML models were selected based on our prior knowledge of previous models that were commonly used in healthcare settings for efficiency. L1-regularized logistic regression is a penalised regression method using the sum of the absolute vector values for regularisation. It has an ability to reduce overfitting and shrink coefficients of unimportant variables to zeros, and therefore, feature selection on high dimensional data is not needed [42, 43]. Decision trees are a flowchart-like structure where each node shows a simple decision rule on a predictor that best separates the outcome into two groups with the most disparate probabilities of event. Random forest is an ensemble of decision trees created by using bootstrap samples of the data and random feature selection in tree induction [44, 45]. Since this split is binary, it is able to capture non-linearities in the data, as multiple splits on the same predictor can occur within one tree. One of the advantages of the Random forest and Xgboost models is that they do not require much effort to tune parameters [46, 47]. Neural networks are designed based on neuronal activation structure of the human brain using synaptic weights that represent 'hidden layers' between inputs and outputs [41]. These methods were applied in this research because they can work with disease risk prediction (i.e., binary data), are commonly used in the literature, and are interpretable (for L1-regularized logistic regression, decision trees and random forest), which is of interest to policy makers. Further discussion on the pros and cons of these methods in health literature has been provided elsewhere [41].

Following Zafar et al. [36], two fold-cross validations were performed on the training data. Parameter tuning was performed using area under the receiver operating curve (AUC) as an evaluation matrix. Models were coded and analyzed in R version 3.3.0. All ML models were trained using the same training datasets and tested on the same test datasets to allow comparison of their predictive power. The main indicator AUC was used to evaluate the predictive performance of the ML models.

To identify potential variables associated with the CVD events among people with diabetes, we used the built-in feature of the random forest model. The random forest's variable of importance was calculated based on a Gini index, which gives a relative ranking to all variables [45, 48]. This Gini index is suitable for the classification problem, which is the prediction outcome of this research [49].

Model optimization

ML models were trained to maximize the AUC indicators, either for the whole NZ population with diabetes or for a particular ethnic group (e.g., Asian) as per Fig. 2. This optimization aimed to get better health outcome predictions by ethnicity to reduce health inequity gaps. Two key steps to improve prediction in a pre-defined ML model are data inputs and parameter tuning. The optimization scenarios A and B used nationwide observations, but tuning models and predicting health outcomes are for, say, Māori only. Optimization C is a common ML model stratifying data inputs and parameters tuning by ethnicity. Our measure is somewhat similar to the group fit measurement employed by McGuire et al. [50] However, while these authors used group fit for the total payment ratio received for groups by health condition (cancer, heart health, diabetes and mental health), our group fit was AUC by ethnicity. Furthermore, McGuire et al. set up a constraint on the group fit measurement (i.e., the total payment ratio equals one reflecting a balance between budgeted and actual health expenditures), but we optimized our group fit level by ethnicity through parameter tuning. Tuning models' parameters were selected with cross-validation, with parameters being set on a minimum-maximum range (i.e., number of trees in random forest, and depth of a tree). A record of the AUC value was kept for each combination of models' parameters, and then the parameter combination with the highest AUC score was chosen. This process was repeated across different optimization scenarios A, B and C. The

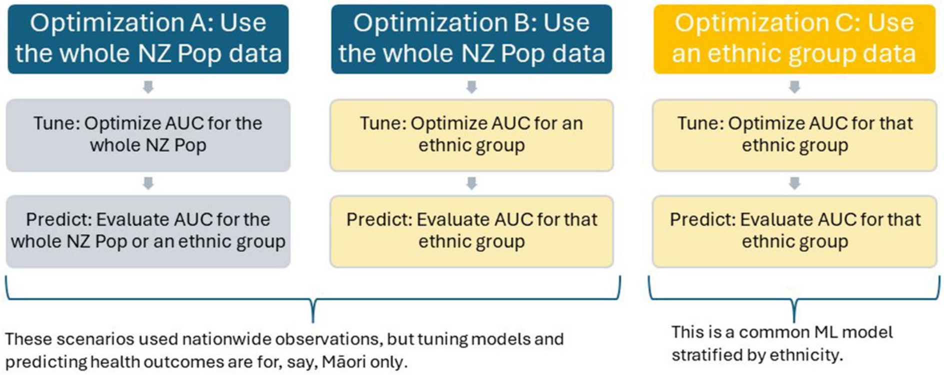
Optimization A: Use the whole NZ Pop data
Tune: Optimize AUC for the whole NZ Pop
Predict: Evaluate AUC for the whole NZ Pop or an ethnic group
Optimization B: Use the whole NZ Pop data
Tune: Optimize AUC for an ethnic group
Predict: Evaluate AUC for that ethnic group
Optimization C: Use an ethnic group data
Tune: Optimize AUC for that ethnic group
Predict: Evaluate AUC for that ethnic group
These scenarios used nationwide observations, but tuning models and predicting health outcomes are for, say, Māori only.
This is a common ML model stratified by ethnicity.

Fig. 2 Optimization scenarios with different data subsets and evaluation indicators for populations with diabetes

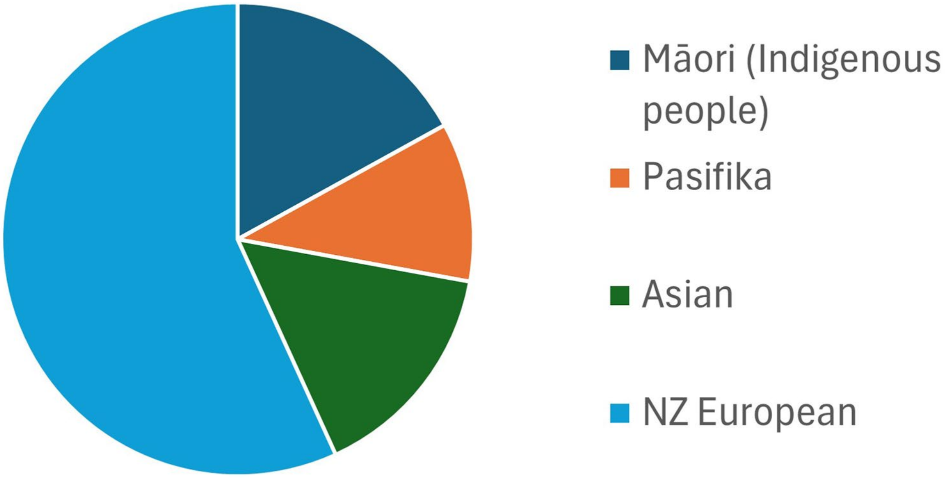

Fig. 3 The ethnic composition of Aotearoa New Zealand in the study population

model with the set of model parameters that predicted the highest value of AUC among these optimization scenarios was selected as the final prediction model.

Results

Descriptive results: The ethnicity composition of the study population in NZ was presented in Fig. 3, with Māori accounting for 17%. As shown in Table 1, there were approximately 145,600 NZ residents with diabetes who were aged 30–74 and with complete data on basic demographic information: age and sex. There were less than 0.5% of observations having missing ethnicity data, and less than 10% missing smoking status data. All observations with missing data other than age and sex were included in the analyses and were implicitly treated as missing data. This means we excluded all observations with missing age and sex data, which were very rare as these were imputed by Stats NZ. For some variables, it was impossible to identify whether there was any missing data, such as a CVD event or medications prescribed. So we performed external validation of our health events with Ministry of Health Burden of Disease report, such as CVD events and diabetes prevalence, and our peer-reviewed disease modelling studies [11, 27, 31, 51–55]. The term "implicitly treated as missing data" means that if there were three deprivation levels, and there was no information for a particular individual on all three deprivation binary variables (i.e., all three zeros instead of a one and two zeros), then that individual's information on deprivation was categorized as "missing". No sensitivity analyses were employed to account for missing data; however, these datasets were large, high-quality and were used for official statistics (e.g., census data, hospital and medications prescribed data) for many years. Table 2; Fig. 4 present CVD incidence rates among people with diabetes by various predictors, in particular age, sex, ethnicity, deprivation decile, smoking status, and employment status.

Models were trained and tested using data for the NZ study population aged 30–74 years with diabetes in 2013–2015 (approximately 74,600 individuals, Table 1), were optimized for the indicator (AUC) for this population, and were predicted by ethnic group. After initial investigation of the models' prediction, we only

Table 1 Descriptions of variables included in the analysis in both study and validation datasets

Study variables	Study dataset: *N* (counts of observations) (% of the observations)	Validation dataset: *N* (%)
Total population aged 30–74 years with diabetes in NZ	74,600 (100%)	71,000 (100%)
Female	36,600 (49.1%)	35,300 (49.7%)
Male	38,000 (50.9%)	35,700 (50.3%)
Māori	12,800 (17.2%)	12,100 (17.1%)
Pasifika	8,200 (11%)	7,980 (11.2%)
Asian	11,600 (15.5%)	11,000 (15.6%)
NZ European	42,800 (57.5%)	40,700 (57.4%)
Māori (mixed ethnicity – up to three)	3,220 (4.3%)	3,030 (4.3%)
Pasifika (mixed ethnicity – up to three)	990 (1.3%)	880 (1.3%)
Asian (mixed ethnicity – up to three	510 (0.7%)	510 (0.7%)
NZ European (mixed ethnicity – up to three)	510 (0.7%)	550 (0.8%)
Mean age (years)	57	57
Deprivation high*	29,300 (39.3%)	28,000 (39.4%)
Deprivation medium*	27,800 (37.2%)	26,000 (36.7%)
Deprivation low*	17,300 (23.3%)	16,800 (23.7%)
Current smokers	10,500 (14.1%)	9,900 (13.9%)
Ex-smokers	21,500 (28.9%)	20,500 (28.9%)
Non-smokers (note: ~7% missing smoking status data)	38,700 (52%)	37,000 (52.1%)
Having a post-graduate qualification	24,900 (33.4%)	24,000 (33.8%)
In-paid employment	39,500 (53%)	38,600 (54.4%)
Having gout	8,500 (11.4%)	7,900 (11.2%)
Having cancer	5,060 (6.8%)	4,830 (6.8%)
Having traumatic brain injury	1,360 (1.8%)	1,200 (1.7%)

Notes Deprivation low is deprivation deciles 1–3, medium is 4–7, and high is 8–10

Table 2 CVD incidence counts and rates during the three-year follow-up period by demographic information in the study population

Study variables	CVD incidence event counts (rates) for the study dataset	CVD incidence event counts (rates) for the validation dataset
Total population aged 30–74 years with diabetes	3,430 (4.8%)	3,090 (4.56%)
Female	2,250 (5.9%)	1,980 (5.55%)
Male	1,190 (3.2%)	1,110 (3.16%)
Māori	640 (5%)	590 (4.9%)
Non-Māori	2,790 (4.5%)	2,510 (4.3%)
Deprivation lowest (two deciles 1–2)	390 (15.5%)	410 (3.7%)
Deprivation low (3–4)	500 (4.1%)	460 (3.9%)
Deprivation medium (5–6)	660 (4.8%)	560 (4.4%)
Deprivation high (7–8)	790 (4.9%)	690 (4.6%)
Deprivation highest (9–10)	1,090 (5.2%)	970 (4.8%)
Current smokers	610 (5.8%)	520 (5.2%)
Not current smokers	2,823 (4.41%)	2,580 (4.22%)
In-paid employment	1,410 (3.6%)	1,460 (3.8%)
Not in-paid employment	2,030 (5.8%)	1,640 (5.1%)
$30 \le \text{age} < 40$ (years)	15 (0.5%)	15 (0.5%)
$40 \le \text{age} < 50$ (years)	170 (1.6%)	170 (1.7%)
$50 \le \text{age} < 60$ (years)	510 (3.4%)	460 (3.2%)
$60 \le \text{age} < 70$ (years)	1,180 (4.8%)	1,160 (4.9%)
$70 \le \text{age} < 80$ (years)	1,550 (7.3%)	1,290 (6.5%)

Notes The final sample was randomly divided into a study and a validation dataset, by ethnicity

proceeded with and reported prediction results from the two best performing models: random forest and Xgboost. These models were also commonly used for dealing with rare data (approximately 5% of people with CVD events out of the total study population over the study period) as per those in this analysis [20, 41]. We used an up-sampling method to increase the number of records with CVD events from about 5–20% and 30% of the total

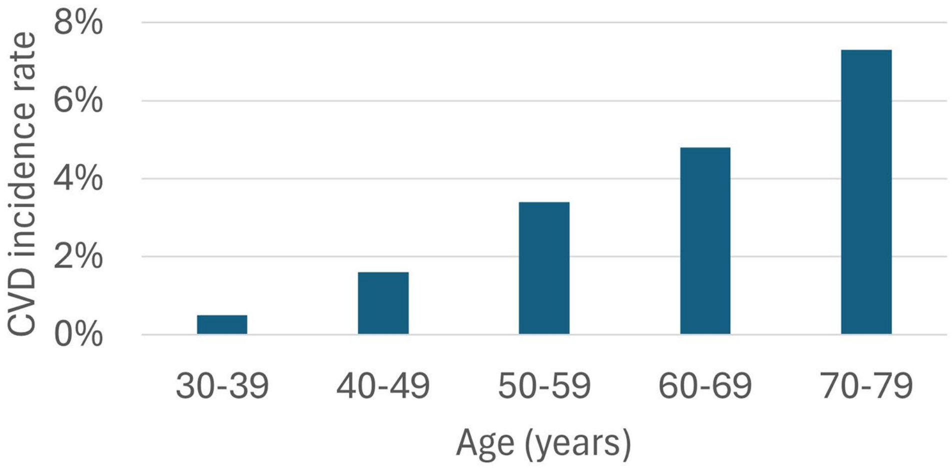

Fig. 4 CVD incidence rates during the three-year follow-up period by age group in the study population

training sets, however, they did not improve our model predictions.

Model performance by ethnicity across subsets of data are presented in Table S3. Models were trained with data for the 2013–2015 period, and results were predicted by ethnic group in the same time period. There were three optimization scenarios as described in Fig. 2. Results suggested that using all data (i.e., all observations for the study population) for training and optimizing all data indicators (Optimization A) improved the prediction compared to using sub-ethnicity data only (Optimization C) by 0.05 AUC (7.0%), 0.04 (5.2%), and 0.03 (4.8%) for Asian, Māori and Pasifika peoples, respectively, using the Xgboost model. Overall, Xgboost models benefited more from using population data than other models. With this current dataset, there were no benefits from optimizing ethnicity indicators (e.g., building the optimal prediction so that it predicts best for Māori). Optimizing ethnicity indicators meant the model performance was assessed based on the prediction of a particular ethnic group. For example, if the Māori ethnicity indicator was optimized, the model's parameters were selected if they produced the best prediction for Māori based on the AUC score. Training data can include observations for all ethnicities, but only observations from Māori individuals were used to calculate the AUC score.

Table S4 shows gaps in model performance by ethnicity for the main indicator (AUC), using the study dataset in the 2013–2015 period. The Xgboost models performed better in term of equity gaps, with an overall prediction improvement of 0.1% on average for sub-ethnic groups compared to the whole NZ population. The average improvement for the random forest (RF) models was −2.3%, that is, the prediction for sub-ethnic groups was not as good as for the whole NZ population with diabetes.

Table 3; Fig. 5 present model performance by ethnicity across two time periods (2013–2015 and 2016–2018). When there was no change in time period (that is training and test datasets were in the same period), results suggested that Xgboost models outperformed all other ML models in term of preventing future CVD events – based on AUC, across ethnicity and time periods. In particular, the average AUC by time period was 0.74 for the whole NZ population with diabetes, and similarly for other populations: Asian (0.74), Māori (0.76), and Pasifika population (0.73). Compared to the RF models, the prediction by Xgboost models was improved by 6.4% (0.74 vs. 0.70) for the whole NZ population with diabetes and 10% (0.73 vs. 0.66) for the Pasifika population.

In terms of predicting future CVD events, the RF models were quite similar to the Xgboost model for the whole NZ population with diabetes. But both models were slightly worse at predicting future events for Māori (absolute AUC gaps: 0.03 for RF and 0.05 for Xgboost, or about 4% and 7% worse, respectively). Both models seem to perform well for the Asian population in predicting future CVD events.

Table 4 presents a list of variables of importance generated by the RF models. Several main traditional risk factors for CVD were picked up (i.e., being given higher ranking) by the RF models, in particular age, deprivation, and the number of years living with diabetes. Other socio-economic factors were also rated highly by the RF model, including: geographical area, income, deprivation, and occupation.

Discussion

Interpretation of the main results

Our study demonstrated the feasibility and value of applying ML methods to administrative health data, including considering fairness in terms of ethnicity. In this case, fairness is more about making the prediction results more accurate for Māori and, hence, potentially facilitating better diagnosis and treatment relative to more accurate prediction for the overall population.

Table 3 Model performance by ethnicity across two time periods (2013–2015 2016–2018) *(models were trained using data for the whole NZ population with diabetes in 2013–2015*, were optimized for the indicator (AUC) for this population and were predicted by ethnic group)

Models*	Model prediction (AUC) for the 2013–2015 period (current period)	Model prediction (AUC) for the 2016–2018 period (future period)	Average AUC for both time periods	Absolute AUC gaps between two time periods (<=0 means at least equal prediction)
Random forest (RF) all ethnic groups	0.70	0.70	**0.70**	**0.00**
RF Asian	0.68	0.71	0.70	-0.03
RF Māori	0.70	0.68	0.69	0.03
RF Pasifika peoples	0.66	0.68	0.67	-0.02
Xgboost all ethnic groups	0.74	0.73	**0.74**	**0.01**
Xgboost Asian	0.74	0.76	0.75	-0.01
Xgboost Māori	0.76	0.71	0.73	0.05
Xgboost Pasifika peoples	0.73	0.71	0.72	0.02

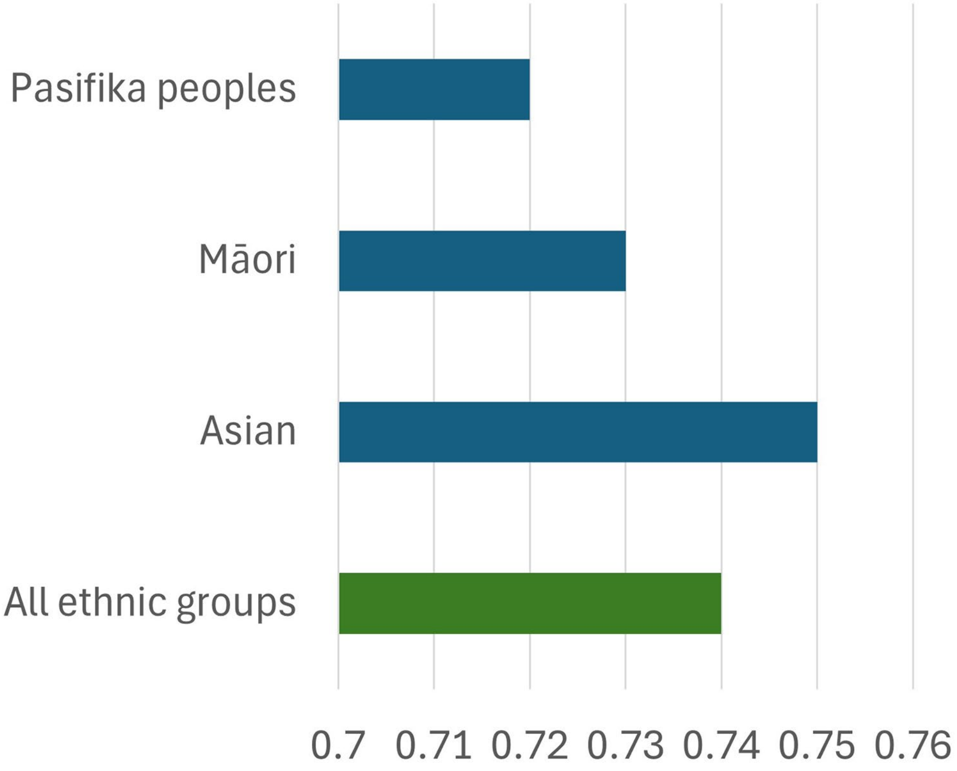

Fig. 5 Xgboost model performance by ethnicity on average over the 2013-2018 period

Table 4 Variables of Importance generated by the random forest model

Rank	Variable of Importance	In a traditional regression model (i.e., the NZ PREDICT equation) [1] (Yes/No)
1	Age	Yes
2	Geographical area (a smallest geographical area in NZ with code by regional council, territorial authority, ward, and area unit)	No
3	Having any hospitalization events	No
4	The number of years living with diabetes	Yes
5	Deprivation level	Yes
6	Having prescribed antiplatelet medicine	Yes
7	Income level	No
8	Having other chronic conditions prior to 2014, including cancer, gout, and traumatic brain injury.	No
9	Occupation (an occupation level that is not classified)	No
10	Having prescribed blood pressure lowering medicine	Yes

Our best model Xgboost can predict the three-year risk of CVD events in those with diabetes with an average AUC of 0.74. Our model was trained on test data for the NZ population with diabetes, which includes marked diversity by demographic and socio-economic variables. Our modelling was also validated in terms of the prediction of future events. There seemed to be no trade-offs between the overall fit of the ML model and the fairness measurement in our analyses. We used an up-sampling method to increase the number of records with CVD events from about 5–10% and 15% of the total training sets. However, they did not improve our model predictions.

The results suggested that the models generally performed slightly better for large population groups. In general, it should not be surprising that for a larger sample, i.e., all ethnicities in this study, the models performed better. However, the better performance of the model depends on both the larger sample size and the less heterogeneity of the population's characteristics. If the sample size is small but the characteristics of the study population are quite homogenous or have less variation, then the model performs better. In addition, if the characteristics of the population are highly correlated to the outcome, even with a small sample size, the model can still perform better than with a larger sample size. Indeed, in Table 3, we showed that the performance of the Xgboost model was better for the Asian population than all ethnic groups combined for the period 2016–2018 and the total period.

Our models performed reasonably well in comparison to the literature. In particular, the model to predict CVD risks among people with diabetes in the NZ setting from the 400,000-person primary care cohort study reported C-statistics of 0.73 and 0.69 for women and men, respectively [1]. This CVD risk prediction model contained clinical variables such as BMI and systolic blood pressure, but it should be highlighted that our models did not have rich clinical features as per the traditional risk prediction models but were still able to produce comparable prediction results. We expected that if these clinical features were incorporated, the performance of the models would be improved. Our model's performance was lower than the one developed in Canada [9] (AUC of 0.74 vs. 0.79) but this Canadian work had more data points. Of note is that, AUC and C-statistics are identical in the case of binary outcome, which is used in this study [56].

Similar to the findings by the study in Canada, our variables of importance also picked up socio-economic factors as important variables in the prediction result [9]. These variables include geographical area, income, occupation, and education level. The Canada study indicated that socio-demographic factors such as length of stay in Canada for immigrants and ethnic concentration in the area of residence, play an important role in model prediction. Our study's contribution was that, in the absence of important clinical data, we showed the value of routinely collected administrative data at a national level that can potentially be used to predict CVD outcomes.

Study strengths and limitations

This study benefited from NZ having established some of the most comprehensive administrative health data holdings in the world, covering nearly the total population due to its universal healthcare system and digital government [11].

Our ML models were validated against future time with no significant differences in model performance. These results were applied at both the total population level and the ethnic group level. Both Random Forest and Xgboost models employed in this study are seen as a “black box” needing external validations and are more useful for prediction rather than causal inference [46, 47], which fits the scope of this study.

Nevertheless, this type of study is not currently easy to perform given that data used are held by the central government and the current computing infrastructure does not easily facilitate developing and running ML models on such large datasets. However, these constraints may ease with the expansion in size and speed of computing systems.

It is important to emphasize that our variable of importance results (Table 4) cannot be interpreted as causal for CVD events, but these novel important factors can be further tested for causal inference such as assessing omitted variable bias, heteroscedasticity or endogeneity issues, which are out of the scope of this research [57]. In our study, the inclusion of these factors did improve the prediction ability, such as having any hospitalization events, which was indeed in line with the literature in disease prediction [58]. Even though our important variables result is not causal, it will enable others to further explore these variables for guiding policies in managing CVD risk factors among people with diabetes.

In Table 4, we implied that having at least one chronic condition of cancer, gout or traumatic brain injury is an important predictor for CVD events. We noted that identifying which comorbidities increase the risk of a CVD event is more useful, and, as described in our *Methods* section, we included asthma, anxiety and depression, chronic kidney disease, and dementia, but these predictors did not top the model's variable of importance list. We tested a model with a number of hospitalizations, but it did not perform as well as the model with any hospitalization events. Having an unclassified occupation may be a proxy for unemployment and, therefore, could potentially be a predictor for higher CVD events. Further research should investigate this predictor in more detail.

Finally, as noted above, our models lack clinical risk factors such as BMI and systolic blood pressure.

Implications for health system

Using ML with administrative health data provide an opportunity for automated analytic approaches to improve the efficiency and effectiveness of primary and secondary health prevention efforts, and address systemic barriers to diabetes care [9]. Our findings suggest that ML can be capitalized to draw insights from administrative social and health data to improve health management and improve health equity.

While risk for CVD events among people with diabetes have been better managed in recent years, they remain a large burden because the incidence of diabetes continues to grow. Thus, there is a need to effectively prevent and manage diabetes complications at not only the individual patient level but also system levels. Health inequities by ethnicity in NZ have long been recognized, but these gaps are still persisted [59–61]. More urgent action and policy interventions both within and beyond the health system are needed to reduce health burdens in marginalized populations [62]. In our prediction tasks, the model might automatically select the optimal set of parameters based on characteristics of the majority population, so that it best fits the model; and therefore, ignoring particular characteristics of the minority population at risk. Hence, by optimizing model performance based on ethnicity, we minimized the risk that the best prediction model would bias towards better prediction for the major ethnic population. However, we found that there was no trade-off between prediction performance and equity for other indicators; that is we can improve model prediction and reduce model performance gaps by ethnicity simultaneously. Furthermore, model training separately by ethnicity did not work well in this setting, so it appears best to use population data with ethnicity information, rather than train separate model for each ethnicity.

Even though our aim was to develop a prediction model for deployment at a population level, our variables of importance can still be further tested (i.e., through a lasso logistic regression or a causal random forest model) [57, 63] to create a checklist to be used in the primary healthcare setting. Linked administrative and health databases typically have millions of records spread across multiple datasets making it highly challenging to work with. Moreover, predictive patterns inferred by the model at this scale can identify new trends or new risk factors at the population level. These variables may not be available in clinical prediction models as they generally exclude such types of features and mainly focus on health data for each patient. Thus the application of a ML model developed on administrative datasets to allocate resources and plan policies at a population level to improve diabetes complications outcomes could offer a data-driven approach to addressing health inequities [9].

Better prediction of disease risk factors can potentially provide evidence and data for policy decision-making to reduce the disease burden and reduce associated health inequities. For example, this type of targeted intervention response has already happened with NZ's CVD guidelines, whereby Māori men are advised to start CVD risk assessment at the age of 30 years, while for men and women without known risk factors the equivalent ages are 45 and 55 years, respectively [28]. In addition, people with severe mental illness are advised to have this CVD assessment at the age of 25 years, which is at least 20 years earlier than individuals without known risk factors. As other risk factors are clarified, then such targeting can be further fine-tuned for different ethnic and socio-economic groups.

Future research

With the improvement in computing power that allows processing a large amount of data, the number of features can be expanded to investigate yet unknown CVD risk factors to target public health or individual-level interventions. The methodology of this study could also be applicable to other chronic diseases in NZ.

Future analysis may benefit from better accounting the possibility of misclassification in terms of ethnicity, such as the misclassification of Māori as a non-Māori [64], in order to better account for equity issues in NZ.

Employing deep learning techniques that are suitable for large datasets with a large set of predictors might improve model prediction. However, it should be noted that this type of model must be built in a confidential data lab managed by the NZ Government. Also, the most common software for building machine learning models among computer scientists, Python, is not currently available in this confidential environment. In addition, limited computing power provided in this environment might prevent deep learning techniques from being employed successfully.

Further research should also look at investigating the impact of different types of data (e.g., clinical or health-related versus socio-economic) on prediction accuracy.

Conclusions

We provide further evidence that ML with administrative health data can be used for meaningful future prediction of health outcomes. As such, it could be utilized to inform health planning and healthcare resource allocation for diabetes management and the prevention of CVD events. Importantly, the ML model performance was only slightly different between ethnic groups in the NZ context and datasets. This may suggest limited scope for developing prediction models by ethnic group and that the major ways to reduce existing inequitable health outcomes is probably via improved delivery of prevention and management to those groups with diabetes at highest need.

Abbreviations

AUC	Area under the Receiver Operating Characteristics curve
CVD	Cardiovascular Diseases
IDI	Integrated Data Infrastructure
ML	Machine Learning
NZ	Aotearoa New Zealand
RF	Random Forest
Stats NZ –	The New Zealand Government agency for official statistic

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02678-x.

Supplementary Material 1

Acknowledgements

Access to the data presented was managed by Statistics NZ but under strict micro-data access protocols and in accordance with the security and confidentiality provisions of the Statistics Act 1975. Our findings are not Official Statistics. The opinions, findings, recommendations, and conclusions expresses are those of the authors, and not Statistics NZ and the University of Otago.

The results are based in part on tax data supplied by Inland Revenue to Statistics NZ under the Tax Administration Act 1994. This tax data must be used only for statistical purposes, and no individual information may be published or disclosed in any other form, or provided to Inland Revenue for administrative or regulatory purposes. Any person who has had access to the unit record data has certified that they have been shown, have read, and have understood Sect. 81 of the Tax Administration Act 1994, which relates to secrecy. Any discussion of data limitations or weaknesses is in the context of using the IDI for statistical purposes, and is not related to the data's ability to support Inland Revenue's core operational requirements.

We thank Sherri Rose for her useful advice on machine learning fairness at early stages of this research.

Author contributions

N.N., N.W., J.K., T.T. designed the study. N.N., T.T. built the ML models. N.N. ran models and produced the results. N.N., N.W., J.K., T.T. interpreted the results. N.N. wrote the first draft. N.W., T.T. contributed significantly to the writing. All authors reviewed and approved the final draft of the manuscript.

Funding

This work was funded by the Royal Society Te Apārangi. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability

Access to the anonymised data used in this study was provided by Stats NZ under the security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular person, household, business, or organisation, and the results in this paper have been confidentialised to protect these groups from identification and to keep their data safe.

Code available upon request but note that data for running this code are only available in a strict confidential environment managed by Stats NZ.

Declarations

Ethics approval and consent to participate

All methods were carried out in accordance with relevant guidelines and regulations. The study and study protocols were approved by University of Otago ethics approval processes, reference number HD20/012. There were no participants directly involved in this study study and it used anonymised data. This is a retrospective study hence informed consent was waived by the University of Otago Ethics Committee.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 31 August 2023 / Accepted: 10 September 2024

Published online: 27 September 2024

References

1. Pylypchuk R, Wells S, Kerr A, Poppe K, Harwood M, Mehta S, et al. Cardiovascular risk prediction in type 2 diabetes before and after widespread screening: a derivation and validation study. Lancet. 2021;397(10291):2264–74.
2. Stanaway JD, Afshin A, Gakidou E, Lim SS, Abate D, Abate KH, et al. Global, regional, and national comparative risk assessment of 84 behavioural, environmental and occupational, and metabolic risks or clusters of risks for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of Disease Study 2017. Lancet. 2018;392(10159):1923–94.
3. Blakely T, Kvizhinadze G, Atkinson J, Dieleman J, Clarke P. Health system costs for individual and comorbid noncommunicable diseases: an analysis of publicly funded health events from New Zealand. PLoS Med. 2019;16(1):e1002716.
4. Nghiem N, Atkinson J, Nguyen BP, Tran-Duy A, Wilson N. Predicting high health-cost users among people with cardiovascular disease using machine learning and nationwide linked social administrative datasets. Health Econ Rev. 2023;13(1):1–13.
5. Ministry of Health. Diabetes – Māori health statistics, https://www.health.govt.nz/our-work/populations/maori-health/tatau-kahukura-maori-health-statistics/nga-mana-hauora-tutohu-health-status-indicators/diabetes 2015 [.
6. Coppell KJ, Mann JI, Williams SM, Jo E, Drury PL, Miller JC, et al. Prevalence of diagnosed and undiagnosed diabetes and prediabetes in New Zealand: findings from the 2008/09 adult Nutrition Survey. NZ Med J. 2013;126(1370):23–42.
7. Gurney J, Stanley J, Sarfati D. The inequity of morbidity: disparities in the prevalence of morbidity between ethnic groups in New Zealand. J Comorbidity. 2020;10:2235042X20971168.
8. Anderson KM, Odell PM, Wilson PWF, Kannel WB. Cardiovascular disease risk profiles. Am Heart J. 1991;121(1, Part 2):293–8.
9. Ravaut M, Sadeghi H, Leung KK, Volkovs M, Kornas K, Harish V, et al. Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data. Npj Digit Med. 2021;4(1):24.
10. Yu D, Zhao Z, Osuagwu UL, Pickering K, Baker J, Cutfield R, et al. Ethnic differences in mortality and hospital admission rates between Māori, Pacific, and European New zealanders with type 2 diabetes between 1994 and 2018: a retrospective, population-based, longitudinal cohort study. Lancet Global Health. 2020;9(2):209–17.
11. Camacho X, Nedkoff L, Wright FL, Nghiem N, Buajitti E, Goldacre R, et al. Relative contribution of trends in myocardial infarction event rates and case fatality to declines in mortality: an international comparative study of 1·95 million events in 80·4 million people in four countries. Lancet Public Health. 2022;7(3):e229–39.
12. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. Eur Heart J. 2016;38(23):1805–14.
13. Crown WH. Potential application of machine learning in health outcomes research and some statistical cautions. Value Health. 2015;18(2):137–40.
14. Hofman JM, Sharma A, Watts DJ. Prediction and explanation in social systems. Science. 2017;355(6324):486–8.
15. Subrahmanian VS, Kumar S. Predicting human behavior: the next frontiers. Science. 2017;355(6324).
16. Narain R, Saxena S, Goyal AK. Cardiovascular risk prediction: a comparative study of Framingham and quantum neural network based approach. Patient Prefer Adherence. 2016;10:1259–70.
17. Tay D, Poh CL, Kitney RI. A novel neural-inspired learning algorithm with application to clinical risk prediction. J Biomed Inf. 2015;54:305–14.
18. Wolfson J, Bandyopadhyay S, Elidrisi M, Vazquez-Benitez G, Vock DM, Musgrove D, et al. A naive Bayes machine learning approach to risk prediction using censored, time-to-event data. Stat Med. 2015;34(21):2941–57.
19. D'Ascenzo F, De Filippo O, Gallone G, Mittone G, Deriu MA, Iannaccone M, et al. Machine learning-based prediction of adverse events following an acute coronary syndrome (PRAISE): a modelling study of pooled datasets. Lancet. 2021;397(10270):199–207.
20. Mullainathan S, Spiess J. Machine learning: an applied econometric approach. J Economic Perspect. 2017;31(2):87–106.
21. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE. 2017;12(4):e0174944.
22. Pylypchuk R, Wells S, Kerr A, Poppe K, Riddell T, Harwood M, et al. Cardiovascular disease risk prediction equations in 400 000 primary care patients in New Zealand: a derivation and validation study. Lancet. 2018;391(10133):1897–907.
23. Mehta S, Jackson R, Pylypchuk R, Poppe K, Wells S, Kerr AJ. Development and validation of alternative cardiovascular risk prediction equations for population health planning: a routine health data linkage study of 1.7 million new zealanders. Int J Epidemiol. 2018;47(5):1571–84.
24. Stats NZ. IDI MOH Chronic Condition/Significant Health Event Cohort data, URL: https://datainfoplus.stats.govt.nz/Item/nz.govt.stats/ac775e86-9f66-486a-adb9-64b0f512c54c 2015 [.
25. Stats NZ. https://www.stats.govt.nz/integrated-data/how-we-keep-integrated-data-safe/. 2019.
26. Ministry of Health. IDI Data Dictionary. Chronic condition/significant health event cohort (November 2015 edition). www.stats.govt.nz. 2015 [.
27. Blakely T, Cobiac LJ, Cleghorn CL, Pearson AL, van der Deen FS, Kvizhinadze G, et al. Health, health inequality, and cost impacts of annual increases in tobacco tax: multistate life table modeling in New Zealand. PLoS Med. 2015;12(7):e1001856.
28. Ministry of Health. Cardiovascular disease risk assessment and management for primary care. Ministry of Health Wellington; 2018.
29. Atkinson J, Salmond C, Crampton P. NZDep2013 index of deprivation. Wellington: Department of Public Health, University of Otago. 2014;5541:1–64.
30. Crampton P, Salmond C, Atkinson J. A comparison of the NZDep and New Zealand IMD indexes of socioeconomic deprivation. Kōtuitui: New Z J Social Sci Online. 2020;15(1):154–69.
31. Ministry of Health. Health Loss in New Zealand 1990–2013. 2016.
32. Ministry of Health. BDS 2016 MoH ways-and-means-final. 2016.
33. Buddeke J, Bots ML, Van Dis I, Visseren FL, Hollander M, Schellevis FG, et al. Comorbidity in patients with cardiovascular disease in primary care: a cohort study with routine healthcare data. Br J Gen Pract. 2019;69(683):e398–406.
34. Tran J, Norton R, Conrad N, Rahimian F, Canoy D, Nazarzadeh M, et al. Patterns and temporal trends of comorbidity among adult patients with incident

cardiovascular disease in the UK between 2000 and 2014: a population-based cohort study. PLoS Med. 2018;15(3):e1002513.

35. Buddeke J, Bots ML, van Dis I, Liem A, Visseren FL, Vaartjes I. Trends in comorbidity in patients hospitalised for cardiovascular disease. Int J Cardiol. 2017;248:382–8.
36. Zafar MB, Valera I, Gomez-Rodriguez M, Gummadi KP. Fairness constraints: a flexible approach for fair classification. J Mach Learn Res. 2019;20(1):2737–78.
37. Kohavi R, editor. A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai; 1995: Montreal, Canada.
38. Mooney SJ, Pejaver V. Big data in public health: terminology, machine learning, and privacy. Annu Rev Public Health. 2018;39:95–112.
39. Rose S. Mortality risk score prediction in an elderly population using machine learning. Am J Epidemiol. 2013;177(5):443–52.
40. Van der Laan MJ, Rose S. Targeted learning: causal inference for observational and experimental data. Springer Science & Business Media; 2011.
41. Kreatsoulas C, Subramanian S. Machine learning in social epidemiology: learning from experience. SSM-population Health. 2018;4:347.
42. Shi J, Yin W, Osher S, Sajda P. A fast hybrid algorithm for large-scale l1-regularized logistic regression. J Mach Learn Res. 2010;11:713–41.
43. Razavian N, Blecker S, Schmidt AM, Smith-McLallen A, Nigam S, Sontag D. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. Big Data. 2015;3(4):277–87.
44. Genuer R, Poggi J-M, Tuleau-Malot C. Variable selection using random forests. Pattern Recognit Lett. 2010;31(14):2225–36.
45. Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics. 2009;10:1–16.
46. Varian HR. Big data: new tricks for econometrics. J Economic Perspect. 2014;28(2):3–28.
47. Doupe P, Faghmous J, Basu S. Machine Learning for Health Services Researchers. Value Health. 2019;22(7):808–15.
48. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
49. Han H, Guo X, Yu H, editors. Variable selection using mean decrease accuracy and mean decrease gini based on random forest. 2016 7th ieee international conference on software engineering and service science (icsess); 2016: IEEE.
50. McGuire TG, Zink AL, Rose S. Improving the performance of risk adjustment systems: constrained regressions, reinsurance, and variable selection. Am J Health Econ. 2021;7(4):497–521.
51. Nghiem N, Wilson N. Potential impact of COVID-19 related unemployment on increased cardiovascular disease in a high-income country: modeling health loss, cost and equity. PLoS ONE. 2021;16(5):e0246053.
52. Nghiem N, Leung W, Cleghorn C, Blakely T, Wilson N. Mass media promotion of a smartphone smoking cessation app: modelled health and cost-saving impacts. BMC Public Health. 2019;19(1):283.
53. Nghiem N, Knight J, Mizdrak A, Blakely T, Wilson N. Preventive pharmacotherapy for cardiovascular disease: a modelling study considering health gain, costs, and cost-effectiveness when stratifying by absolute risk. Sci Rep. 2019;9(1):19562.
54. Nghiem N, Cleghorn CL, Leung W, Nair N, Deen FSV, Blakely T, et al. A national quitline service and its promotion in the mass media: modelling the health gain, health equity and cost-utility. Tob Control. 2018;27(4):434–41.
55. Nghiem N, Blakely T, Cobiac LJ, Cleghorn CL, Wilson N. The health gains and cost savings of dietary salt reduction interventions, with equity and age distributional aspects. BMC Public Health. 2016;16(1):423.
56. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiol (Cambridge Mass). 2010;21(1):128.
57. Athey S, Imbens GW. The state of Applied Econometrics: causality and policy evaluation. J Economic Perspect. 2017;31(2):3–32.
58. Nghiem N, Atkinson J, Nguyen BP, Tran-Duy A, Wilson N. Predicting high health-cost users among people with cardiovascular disease using machine learning and nationwide linked social administrative datasets. Health Econ Rev. 2023;13(1):9.
59. Nghiem N, Teng A, Cleghorn C, McKerchar C, Wilson N. Using household economic survey data to assess food expenditure patterns and trends in a high-income country with notable health inequities. Sci Rep. 2022;12(1):21703.
60. Nghiem N, Leung W, Doan T. Health promoting and demoting consumption: what accounts for budget share differentials by ethnicity in New Zealand. SSM-Population Health. 2022;19:101204.
61. Wilson N, Cleghorn C, Nghiem N, Blakely T. Prioritization of intervention domains to prevent cardiovascular disease: a country-level case study using global burden of disease and local data. Popul Health Metrics. 2023;21(1):1.
62. Yu D, Zhao Z, Osuagwu UL, Pickering K, Baker J, Cutfield R, et al. Ethnic differences in mortality and hospital admission rates between Māori, Pacific, and European New zealanders with type 2 diabetes between 1994 and 2018: a retrospective, population-based, longitudinal cohort study. Lancet Global Health. 2021;9(2):e209–17.
63. Wager S, Athey S. Estimation and inference of Heterogeneous Treatment effects using Random forests. J Am Stat Assoc. 2018;113(523):1228–42.
64. Shaw C, Atkinson J, Blakely T. (Mis) classification of ethnicity on the New Zealand cancer registry: 1981–2004. New Z Med J (Online). 2009;122(1294).

Publisher's note